1

SEC[illegible] [illegible]IFICATION OF THIS PAGE

JMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a AD-A208 463

1b RESTRICTIVE MARKINGS
NONE

2a

3. DISTRIBUTION/AVAILABILITY OF REPORT
APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFIT/CI/CIA-88-212

6a. NAME OF PERFORMING ORGANIZATION
AFIT STUDENT AT St Mary's University

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
AFIT/CIA

6c. ADDRESS (City, State, and ZIP Code)

7b. ADDRESS (City, State, and ZIP Code)
Wright-Patterson AFB OH 45433-6583

8a. NAME OF FUNDING/SPONSORING ORGANIZATION

8b. OFFICE SYMBOL (If applicable)

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS (City, State, and ZIP Code)

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification) (UNCLASSIFIED)
Meta-Analysis of Armed Service Vocational Aptitude Battery Subtest Validity Data

12. PERSONAL AUTHOR(S)
Nicole S. Stermer

13a. TYPE OF REPORT
THESIS/~~DISSERTATION~~

13b. TIME COVERED
FROM ________ TO ______

14. DATE OF REPORT (Year, Month, Day)
1988

15. PAGE COUNT
68

16. SUPPLEMENTARY NOTATION
APPROVED FOR PUBLIC RELEASE IAW AFR 190-1
ERNEST A. HAYGOOD, 1st Lt, USAF
Executive Officer, Civilian Institution Programs

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

89 6 02 011

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
ERNEST A. HAYGOOD, 1st Lt, USAF

22b. TELEPHONE (Include Area Code)
(513) 255-2259

22c. OFFICE SYMBOL
AFIT/CI

DD Form 1473, JUN 86

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

AFIT/CI "OVERPRINT"

META-ANALYSIS OF
ARMED SERVICE VOCATIONAL APTITUDE BATTERY
SUBTEST VALIDITY DATA

APPROVED:

____
(Supervising Professor)

____

____

APPROVED:

____
(Dean of Graduate School)

DATE:________________

META-ANALYSIS OF

ARMED SERVICE VOCATIONAL APTITUDE BATTERY

SUBTEST VALIDITY DATA

A
THESIS

Presented to the Faculty of the Graduate School
of St. Mary's University in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE

in

Industrial Psychology

BY

Nicole S. Stermer, BA

San Antonio, Texas

December, 1988

Acknowledgements

META-ANALYSIS OF
ARMED SERVICE VOCATIONAL APTITUDE BATTERY
SUBTEST VALIDITY DATA

Nicole S. Stermer

St. Mary's University, 1988

Supervising Professor: Dr. Malcolm J. Ree

This study was conducted to investigate the efficacy of selection procedures associated with the Armed Services Vocational Aptitude Battery (ASVAB). The main hypothesis tested was whether the Armed Forces Qualifying Test (AFQT), an ASVAB subtest composite, is a valid predictor of training success. Subhypotheses investigated whether the AFQT is a more valid predictor of training success than the individual career-specific selector composites. A final hypothesis dealt with the expectation that the AFQT would show a larger improvement in validities for women than for men. It was expected that the AFQT would demonstrate validities at least as high as the selector composites, due to the moderating effects of general cognitive ability. Keywords: Aptitude tests, personnel selection, Gender differences, Theses. (SI [illegible]

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE

CHAPTER 1

INTRODUCTION

As a personnel selection instrument, the Armed Services Vocational Aptitude Battery (ASVAB) is a vital component in the maintenance of a quality military force. By investigating the method in which ASVAB subtest scores are applied in selection procedures, this study attempts to aid the perpetual research efforts to maintain the ASVAB as a state-of-the-art selection instrument.

To elucidate the scope of this research problem, a number of issues involved in personnel selection will be discussed, including the concept of g, and gender differences in intelligence measures. The ASVAB will then be described in terms of its history, content, reliability, and validity. Lastly, validity generalization and meta-analysis will be discussed, as a basis for this study.

History of Personnel Selection

The issues surrounding the topic of personnel selection are much more complex than would initially be assumed by a casual observer. A historical review is required, before delving into this many-faceted topic.

The first group test used to assess abilities particularly for the purpose of personnel selection was developed during World

War I, by a team of psychologists headed by Robert Yerkes (Graham and Lilly, 1984). The resulting Army Alpha was the first multiple choice, objectively scorable, group administerable test devised. As such, it provided the military with a highly efficient means of measuring cognitive ability, or intelligence, of World War I recruits. Though the Army Alpha was developed to reflect a measure of general intelligence, it did contain specific subtests (specifically: Oral Directions, Arithmetrical Reasoning, Practical Judgement, Synonym-Antonym, Disarranged Sentences, Number Series Completion, Analogies, and Information).

At that time in the history of psychological testing, "the prevailing opinion was that man posesses a single, all-important intellectual ability" (Ghiselli, 1973). Charles Spearman, (1904, 1927) was a leading pioneer in the investigation of the concept of intellectual ability. Through his invention of the procedure of factor analysis, Spearman found evidence for his Two-Factor Theory of intelligence. His factor analytic procedures identified intercorrelations among tests of intellectual ability. He defined the commonality among the tests as representing general intelligence, labeled g, which composed the first factor of his theory. The second factor, s, represented a factor that was specific to each test or type of test.

Faith in the existence of Spearman's g faded following Thurstone's (1938) application of his centroid factor method.

Thurstone did not locate g, or any common factor, in a battery of fifty-six tests. Upon this evidence, Thurstone developed his theory of Primary Mental Abilities, which claimed that seven unique factors exist, each representing a separate mental ability, and that no single index could satisfactorily explain the concept of intellectual ability.

Thurstone's rejection of the existence of g did not stand for long. Spearman reviewed Thurstone's procedures, and by reworking the data, did locate the general factor. In addition, when Thurstone attempted to devise tests to measure his primary factors, he found that the primaries were intercorrelated, rather than independent as he had hypothesized (McNemar, 1964).

Though Thurstone devoted many years to the task of searching for pure measures of distinct abilities, he was unable to produce tests which remained uncorrelated with one another when based on a large, representative population (Jensen, 1986). This phenomenon, known as Positive Manifold, contends that mental ability tests with adequate reliabilities, when administered to representative populations, will always result in positive intercorrelations. Thurstone finally admitted that a general factor was required to explain the intercorrelations between his primary factors (Thurstone & Thurstone, 1941).

Despite the preliminary evidence in support of Spearman's g, the general movement in test development began to lean toward measurement of specific abilities, downplaying the concept of g.

Between World Wars I and II, psychologists hypothesized that batteries of tests measuring specific aptitudes could produce a composite score that would be equivalent to an aptitude test tailored to a particular job or curriculum. Hunter (1984) states this hypothesis was not tested until recently, and fails in all but a few specific cases. Though untested, the hypothesis of differential aptitude testing became particularly appealing to the military, due to rapid advances in military technology. During World War II, many leading psychometricians applied multiple regression strategies to aptitude batteries in order to optimally predict performance in technical military career fields. Multiple or differential aptitude batteries were employed in the civilian sector as well.

Validation results of such batteries have led to an emphasis on quantitative and verbal aptitudes, and also on technical aptitudes particularly for military batteries. The three most commonly used differential aptitude batteries at present are: the ASVAB, developed by the Department of Defense (DoD), (U.S. DoD, 1984); the General Aptitude Test Battery (GATB) developed by the U.S. Employment Service (USES, 1970); and the Differential Aptitude Test (DAT) published by The Psychological Corporation.

The ASVAB

The ASVAB was developed in the 1960's for the purposes of selection and classification of United States military personnel. Previously, in the 1950's, the Army, Navy, Marines, and Air Force used test batteries that were developed separately by each service.

The Selective Service Act of 1948 led to development of the Armed Forced Qualification Test (AFQT), the goal of which was to promote a more equitable distribution of abilities among the various services. The AFQT was developed in a joint-services project, for administration to all military applicants. Standardization of the AFQT against the Army's entrance exam, the Army General Classification Test (AGCT) utilized test scores of all military enlisted personnel as of 31 December, 1944. This group became a reference population for subsequent AFQT versions. The AFQT has since been used to compare the distribution of abilities among the services, and to screen applicants for denial of enlistment.

In 1958, the Airman Qualifying Examination (AQE) was introduced into high schools by the Air Force for the purpose of assisting in occupational counseling. Then in 1966, DoD initiated development of a single test to be used for selection and classification by all services. This joint-services project resulted in the ASVAB. Form 1 of the ASVAB replaced the AQE for high school use in 1968, and it was replaced by alternate form 2 in 1973. Alternate form 3 was used for Air Force recruiting beginning in 1973, and by the Marines in 1975.

Early successes of the ASVAB led to its intended use as the sole instrument for selection, classification, and assignment of all service personnel. In 1976, ASVAB form 5 was provided to high schools, while parallel forms 6 and 7 were provided to Military entrance Processing Stations (MEPSs), thus replacing the batteries used by the individual services.

The Air Force Human Resources Laboratory (AFHRL) is currently the lead laboratory responsible for continued ASVAB development and validation. ASVAB forms 8, 9 and 10, which were developed to provide more accurate measures at lower levels of ability, were placed in use in October, 1980. Replacement forms 11, 12, 13 and 14, developed parallel to forms 8, 9 and 10, began use in October, 1984, and are the forms currently used.

The 1980 Profile of American Youth, a DoD-sponsored study, provided a more current reference population against which ASVAB scores could be interpreted. A nationally representative sample of about 12,000 men and women, aged 16 to 23, took form 8AX (parallel to form 8A) of the ASVAB. Bock and Mislevy (1981) provide supporting testimony as to the representativeness of the sample, and the accuracy of the test and procedures used.

The content of the ASVAB forms 8 through 14 consists of

10 subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). The subject areas have been chosen according to their abilities to predict success in military training courses.

The current AFQT used by all services consists of a composite of four ASVAB subtests (AR, WK, PC, and a half-weighting of NO; see Table 1). The AFQT is used for initial selection or rejection of applicants. Each of the services employs additional subtest configurations for classification purposes. The Air Force utilizes four such composites: Mechanical, Administrative, General and Electronic. (See Table 1 for compositions of these composites). The composites, or aptitude indexes (AIs) were developed through multiple regression procedures.

In the recruitment process, an applicant is administered a current form of the ASVAB. His/her AFQT score, if sufficiently high, qualifies the individual for enlistment. The AI scores are then used, in addition to manning requirements and the enlistee's preferences, to place the individual in a specific job, or career field. Some highly selective Air Force career fields employ higher AI cutoff scores than others.

Studies of ASVAB reliabilities have yielded results suggesting

Table 1

ASVAB Subtest Composites

ASVAB Subtest	AFQT	M[1]	A[2]	G[3]	E[4]
General Science (GS)		X			X
Arithmetic Reasoning (AR)	X			X	X
Word Knowledge (WK)	X		X	X	
Paragraph Comprehension (PC)	X		X	X	
Numerical Operations (NO)[5,6]	X		X		
Coding Speed (CS)[5]			X		
Auto and Shop Information (AS)		X			
Mathematics Knowledge (MK)					X
Mechanical Comprehension (MC)		X			
Electronics Information (EI)					X

[1] M = Mechanical Composite

[2] A = Administrative Composite

[3] G = General Composite

[4] E = Electronic Composite

[5] NO and CS are speeded subtests. All others are power subtests.

[6] NO is weighted at 1/2 for AFQT inclusion.

the ASVAB subtests have satisfactory reliability. Ree, Mullins, Mathews and Massey (1982) computed internal consistency reliabilities for the eight power subtests for ASVAB forms 8a, 8b, 9a, 9b, 10a and 10b. The average subtest reliabilities across all forms ranged from .81 to .92. Reliabilities of the two speeded subtests (NO and CS) were indirectly inferred from subtest intercorrelations, which were observed to range from .53 to .70 (U.S. DoD, 1984).

Hunter, Crosson, and Friedman (1985) report reliabilities for all 10 subtests as computed by extrapolation from previous data (Friedman, Streicher, Wing, Grafton & Mitchell, 1983; Kass, Mitchell, Grafton & Wing, 1982). Hunter and his associates utilized a reliability theory formula which forecasts estimates from one population to another which has a different variance, thereby attempting to show what the ASVAB form 8, 9 and 10 subtest correlations should have been, were there no error of measurement. Extrapolated reliabilities thus computed ranged from .66 to .85.

Although internal consistency reliabilities provide useful information, parallel-forms (or alternate-forms) reliabilities are preferred. Parallel-forms methods correspond to the classic definition of reliability as the ratio of true-score variance to observed-score variance. Such methods avoid the possible bias caused by time effects, as encountered in test-retest methods. Parallel-forms methods also may be used with quickly-paced tests (such as the two speeded subtests of the ASVAB, NO and CS), whereas internal consistency methods will overestimate reliabilities of such tests.

Palmer, Hartke, Ree, Welsh and Valentine (1988) computed parallel-forms reliabilities for subtests and composites of ASVAB forms 8, 9, 10 and 11. Their results, based on data from a sample of 75,000 armed service applicants, indicate that for each subtest, reliabilities are similar across all forms. The coefficients observed for the subtests, across all forms, ranged from .67 to .88. The composite coefficients (including the AIs and the AFQT) ranged from .87 to .93. Reliability coefficients were observed to be higher in general for males and whites than for females and blacks or hispanics, although the reliabilities for females, blacks, and hispanics remained adequate. Reliability coefficients for females ranged from .56 to .92; for blacks, from .83 tc .90; for hispanics, from .80 to .90.

This compilation of studies indicates stable, satisfactory levels of reliability across ASVAB forms and subtest composites.

The validity of the ASVAB is the estimation of how useful it is as a tool to predict job performance in military career fields. Job performance measures have not, however, proven to be appropriate criteria for estimating ASVAB validities, since there are no measures available that can be commonly applied across all military occupations.

ASVAB validity coefficients are computed by correlating the test scores with training school performance measures. Training school grades provide objective measures, and are obtainable across all occupations. Furthermore, the content of the training programs is established according to job performance requirements. Within the Air Force, occupational analyses based on the Instructional Systems Development (ISD) process are performed for each career field. These analyses utilize the Comprehensive Occupational Data Analysis Program (CODAP) to ensure that training school content is based upon required job knowledge. Hunter (1984) has stated that job knowledge and job performance measures are very highly correlated:

> Cognitive ability proved to be a very good predictor of job knowledge and is, thus, a good predictor of job performance. To a lesser extent, cognitive ability also predicts job performance directly . . . . data using work-sample tests show that there is a very high correlation between job knowledge and job performance. People who do not know much about the job will perform poorly. The multiple regression of work sample performance onto ability and knowledge shows that it is job knowledge which has the larger direct impact on performance (p 52 and 53).

Thus not only do job knowledge measures accurately predict job performance, but cognitive ability measures, such as technical school grades, are highly appropriate for assessing job knowledge.

Each of the military services conducted validation studies for ASVAB forms 8, 9 and 10, in 1983. Detailed results and conclusions from these studies are presented in the ASVAB Test Manual (U.S. DoD, 1984). For all services, reported validities are sufficiently strong to predict training success. For Air Force data, no major validity differences were seen between black and white or male and female subgroups, for the specialties which had adequate-sized samples and for which validities reached useful levels.

Numerous extraneous variables can effect the measurement of the validity of occupational tests. Determination of ASVAB validities is complicated by the effects of range restriction in both the lower and upper score ranges. Applicants in the low end of the scale are eliminated from consideration, and thus from the validation sample. Likewise, those whose scores fall above certain cutoff scores may be selected for certain exclusive career fields, thereby being excluded from a given validation sample. Corrections for range restriction may be applied, and will be discussed in conjunction with meta-analytic procedures later in this report.

Additional variables which may effect ASVAB validity data are the differences among training course difficulty, course content, and grading systems. Courses can vary from weeks to months in length, and from basic to highly technical skills. Impacting on these variables is the number of students per course. Some longer, more technical courses require a lengthy time period to develop an adequate-sized sample for validation purposes. Also, variations in grading systems exist, such as pass/fail systems versus letter grades, which may impact on validity measures.

Personnel Selection

Employee selection procedures, including recruiting, interviewing, testing and validating selection criteria cost organizations billions of dollars per year. Choice of selection methods, and validation of those methods are thus of vital concern to employers. In addition to ensuring selection of quality personnel, employers must ensure their selection procedures are within the requirements of federal guidelines (Uniform Guidelines on Employee Selection Procedures, 1978). The guidelines are provided with the intention of promoting equal employment opportunities among all persons, but specifically for certain protected groups. The guidelines require validation studies for each occupation for which a selection instrument is used, if that instrument adversely effects hiring of the specified protected groups.

While attempting to avoid adverse impact regarding the hiring of minorities, the guidelines urge employers to seek selection procedures alternative to tests, but that are equally as valid. That standardized tests are valid methods of personnel selection has been sufficiently supported by Ghiselli (1966, 1973), who reviewed hundreds of criterion-related validity studies. He concluded that "for every job, there is at least one type of test which has at least moderate validity" (Ghiselli, 1973, p 477-478). Attempts to validate alternative selection methods, however, have not met with such success. Reilly and Chao (1982) reviewed research on the validities of eight categories of selection methods: 1. biographical data (biodata), 2. interviews, 3. peer evaluation, 4. self-assessments, 5. reference checks, 6. academic performance measures, 7. expert judgement, and 8. projective techniques. They concluded:

> Only biodata and peer evaluation were supported as having validities substantially equal to those for standardized tests . . . . data, where available, offered no clear indication that any of the alternatives met the criterion of having equal validity with less adverse impact (Reilly & Chao, 1982, p 1).

Hunter and Hunter (1984) assessed validities of various predictors, with training performance determined through supervisor ratings (N ranged from 1,789 to 32,124). Ability composites provided the highest validity, of .53, followed by job tryouts at .44, and biodata at .37. The remaining eight predictors ranged in validity from -.01 to .26.

The Uniform Guidelines' test validation requirements are based on the belief that a test which has demonstrated validity in one situation or group will not necessarily have acceptable validity in others. This exemplifies the theory of situational specificity, which contends that although jobs may appear very similar, there are subtle, yet important differences in jobs, tasks, or individuals which moderate the predictor-criterion relationship. For example, a test found to adequately predict performance of Chicago policemen might, due to racial or geographic differences, be invalid for San Francisco policemen.

Prevalence of the theory of situational specificity resulted in numerous validation studies for small groups, and for nearly identical settings. The validity generalization work of Schmidt and Hunter and their associates has shown this theory to be false (Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman & Shane, 1979; Schmidt, Hunter & Urry, 1976). They contend that observed differences in validities across similar situations and jobs are almost entirely due to statistical errors. Their results have consistently indicated that test validities are highly transportable across groups and situations. The concept of validity generalization will be further expanded later in this paper.

Societal consequences of the importance of intelligence and the ability to measure it have definite implications regarding personnel selection procedures. This discussion has shown that a g factor exists, that it is relevant to job performance, and that it can be assessed by use of standardized tests. What, then, are the societal implications of utilizing tests as selection techniques? Two major theories, Functionalist and Revisionist, have argued that differences in intelligence are of little or no importance in the work place.

The Functionalist theorists assume that education provides job-relevant knowledge required for diverse occupations. They believe also that differences in job performance result primarily from specific skills that have been learned, not from a general ability, and that access to education is not equally obtainable for all individuals. Functionalist reforms would focus on revising educational policies by providing more equal access for underprivelaged groups (Gottfredson, 1986).

Revisionist beliefs are similar, yet are more extreme. They contend that intelligence and educational differences are engineered so as to maintain a hierarchical class society. They argue that an occupational hierarchy is unnecessary, and promotes unjustifiable differences in rewards. According to Gottfredson (1986):

> The revisionist perspective argues, moreover, that the

> high differentiation and specialization of work activities in our society, as well as the accompanying large differences in the entrance requirements and socio-economic rewards among occupations, are the result of capitalists intentionally fragmenting and "de-skilling" work in order to increase not the efficiency of work, but their control over workers (p 382).

The solution to the problem, as the Revisionists see it, is to restructure the jobs themselves, or the personnel selection procedures, rather than revise education or training. They assume the outcome to be both increased equality and productivity.

These theories fail to consider the most basic concern of the hiring organizations, namely, the need for obtaining qualified personnel. It is highly unrealistic to expect employers to be able to compensate each employee according to the employee's exact worth. During the hiring procedure employers can only assume, at best, a level of productivity, or worth for specific individuals. Through trial-and-error procedures, organizations have come to identify indicators of applicants' potentials. Educational credentials have come to be an extensively used indicator of success mainly due to their moderately high correlation with intelligence measures (.6)

(Gottfredson, 1986).

Attempts at social revisions with the dual goals of increasing both equality and productivity have met with little success. A trade-off between the two is inevitable. Attempts to increase equality by randomly assigning workers to differentially g-loaded jobs would result in decreased productivity. Conversely, increasing the validity of job-person matching, though perceived fair and equitable, would further exaggerate the present occupational hierarchy (Gottfredson, 1986).

A popular criticism of selection testing is that it promulgates racial and ethnic inequalities, by identifying as more qualified the individuals who are white, or are from middle class or higher socio-economic backgrounds. Major supporters of this belief have been Ralph Nader and Allen Nairn (Nairn, 1980). Nader's raid on the testing industry focused primarily on the Scholastic Aptitude Test (SAT), a college-entrance exam produced by the Educational Testing Service (ETS). Racial and ethnic criticisms against the use of standardized g-loaded tests have simply not held up. Research on this question has consistently revealed that ethnic groups' score differences on ability tests represent true, consistent differences between groups. Spearman (1927) reported marked differences in mean black and white scores on a battery of 10 highly g-loaded tests. Jensen (1985) corroborated Spearman's findings, and reported that no studies had been found to contradict them. The diverging distributions of g among races will likely result in a proportionately lower number of blacks selected than whites, via cognitive ability measures.

That test results may produce ethnic disparity in numbers hired does not lessen the strength of the evidence for g, nor does it weaken the validity of the tests. The theory of differential validity, which contends that validities are different for various groups of people, has been repeatedly disconfirmed (Bartlett, Bobko, Mosier & Hannon, 1978; Hunter, Schmidt & Hunter, 1979; Schmidt, Pearlman & Hunter, 1980). The model of test unfairness adopted in the Uniform Guidelines on Employee Selection Procedures (1978) is the regression model, which defines a test as unfair if it predicts lower levels of job performance for a minority group than that group actually achieves. As stated by Schmidt and Hunter (1981):

> The accumulated evidence on this theory is clear: Lower test scores among minorities are accompanied by lower job performance, exactly as in the case of the majority (p 1131).

We have seen evidence for the diverging distributions of general cognitive ability for blacks versus whites. What about men versus women--could we expect similar distributions between these groups?

Literature reviews (Jensen, 1980; Maccoby, 1966; Maccoby & Jacklin, 1974) have identified three areas in which gender differences appear: verbal, quantitative, and spatial abilities.

Girls appear to verbally outperform boys beginning at very early ages, perhaps even with their first words (McCarthy, 1954). The differences are small, and boys begin to catch up until about age 10 or 11. After puberty however,

> girls average close to a quarter of a standard deviation higher than boys on verbal tasks . . . . The sex difference in verbal ability after puberty appears to be a genuine phenomenon and not just a measurement artifact (Jensen, 1980, p625).

Jensen's review of studies published since 1966 showed that, of 58 studies investigating general intelligence, 15 significantly favored females, whereas only 3 favored males. This is to be expected, in light of the knowledge that many general intelligence tests are heavily weighted on verbal ability, and require reading.

The pattern of quantitative ability is exactly the opposite of verbal ability. Boys are clearly superior as adolescents, and this difference remains throughout life. Differences of one-fifth to two-thirds of a standard deviation have been observed by the end of high school. This is again a true difference, and not an artifact of test bias. After equating males and females on the number of math courses taken, males still emerge superior, suggesting that the difference is not due to greater training for males (Maccoby & Jacklin, 1974).

The third area of observed differences, spatial ability, has been difficult to define and investigate. Smith's (1964) definition was "the perception, retention, andrecognition (or reproduction) of a figure or pattern in its correct proportions" (p 96). Numerous skills have been proposed as relative to spatial ability, including: auditory localization, size-distance constancy, tactual recognition of objects as they change in space, and matching pairs of items that are mirror-reversals or have been rotated in space. Despite the difficulties encountered in clarifying this concept, researchers have consistently indicated that males are superior on spatial abilities, particularly after adolescence. Maccoby's (1966) review showed an advantage for boys on the DAT and the Primary Mental Abilities Test. Jensen (1980) states that only about one-fourth of the females exceed the male median on spatial-visualization tests. The reason for this difference is still unclear. Proposed causal factors include maturation rates, hormonal factors, sex-linkage, and cultural effects.

In general, females perform better than males on verbal tasks, and somewhat better than males on general intelligence tests, while males are superior on quantitative and spatial tasks.

As in the case of racial differences, the gender differences discovered do not adversely effect interpretations of test validities. Differential predictive validity of tests by sex has concentrated on prediction of college grades. In Jensen's (1980) review, he located no studies reporting lower validities for women. For the ASVAB, as previously mentioned, no significant differences were observed between validity data of males and females.

Another consideration regarding personnel selection is the cost. As stated previously, billions of dollars are spent for hiring programs each year. According to Hunter and Schmidt (1982), most major corporations abandoned the use of selection tests during the previous 10 years, in order to comply more fully with federal guidelines. During this same period, the rate of growth in productivity declined from three and a half percent per year, to zero and even below. The authors attributed this decline, at least in part, to decreased test usage. After dramatically lowering their testing standards to require only minimum scores (at approximately a seventh grade level), U.S. Steel's detailed training records showed that:

1. Scores on mastery tests given during training declined markedly;
2. the flunk-out and drop-out rates increased dramatically;
3. average training time and training costs for those who did make it through the program increased

substantially, and

4. average ratings of later performance on the job declined. (Hunter & Schmidt, 1982, p 298).

Schmidt, Hunter, Outerbridge & Trattner (1986) have empirically estimated the economic impact of various selection methods for the federal work force. They determined that, for the federal government, a one-year cohort based on valid, cognitive ability measures, would increase productivity values up to six hundred million dollars for each year the employees are retained. Hunter and Schmidt (1983) estimate a labor savings of 15 to 20 percent for any organization, due solely to selection based on cognitive ability.

Without doubt, the legal, societal and financial aspects of personnel selection must be of vital concern to any organization. According to John Hawk (1986(, employees who are selected on the basis of validity generalization (based on the concept of g) are more productive, learn faster, and are more quality conscious, and as well are more satisfied with their jobs thus resulting in lower turnover rates. Ability tests remain the most valid personnel selection procedure known. The future of personnel selection procedures appears to lie in the direction of improving testing programs, rather than in the search for alternatives to testing.

Validity Generalization

Schmidt and Hunter and their associates have extensively investigated the feasibility of transporting test validities across groups and situations. Such validity generalization research counters the claims of the situational specificity hypothesis. By generalizing, or "borrowing" validities, researchers have shown that numerous, costly validation studies are unnecessary.

The concept of generalizing results across studies or situations is not new. For many years scientists have been combining results from numerous independent studies in order to infer a general conclusion. Integrative reviews have employed various methods, including vote counting (counting of positive and negative results), counting significant results, and averaging statistics across studies. Rosenthal (1978) describes a number of statistical combination procedures.

As a vanguard of the validity generalization movement, Glass (1976, 1977) devised a statistical method he entitled meta-analysis. His method combines the results of individual studies by converting the results into a common metric, coding various characteristics of the studies, then using conventional statistics to determine whether there is an overall effect.

Schmidt and Hunter and their associates have built a meta-analysis procedure based on the ideas of Glass, but incorporated some modifications. The Schmidt-Hunter method, utilizing Bayesian statistics, improves on Glass's procedure by providing corrections for sources of error. Schmidt, Hunter, Pearlman & Shane (1979) identified seven artifactual sources of variance:

1. Differences between studies in criterion reliability;
2. differences between studies in test reliability;
3. differences between studies in range restriction;
4. sampling error (variance due to $N < \infty$);
5. differences between studies in amount and kind of criterion contamination and deficiency;
6. computational and typographical error, and
7. slight differences in factor structure between tests of a given type. (p 260-261).

Schmidt, et al., do not correct for the last three sources of error, having determined that it is difficult to estimate their frequency or magnitude. Not correcting for these error sources results in conservative variance estimates. They also apply a conservative decision rule, which accounts for the fact that only four of the seven sources of artifactual variance are corrected for. According to their rule, they reject the situational specificity hypothesis if 75 percent or more of the variance of the validity coefficients is accounted for by the four corrected artifacts. If a large proportion of observed variability is attributable to artifacts, the conclusion is that the true population variability is negligible, and validity coefficients can be generalized across situations.

A central concept in the Schmidt-Hunter methodology is that validation attempts have historically relied on insufficient sample sizes, and have thus led to erroneous conclusions (Schmidt & Hunter, 1978). Many supposed moderators of predictor-criterion relationships, including race, ethnicity, sex, age, socio-economic status, leadership style, and geographic area, have been identified through belief in the law of small numbers. This law proposes that a small random sample can be considered as representative of a population as a large random sample. The error of this assumption can be shown using an example of race as a moderator. For validation studies, minority samples have been generally smaller than for the maiority. In single group studies, this produced a large number of white-significant, black-nonsignificant findings. Schmidt and Hunter state that the differential validity approach is sound, but requires an extremely large sample size in each group (in the hundreds) to have a .90 probability of detecting true differences (1978). The small sample size problem encountered in single-group validity studies and differential validity approaches are common among job-test validations, since research must often be based on the number of available workers. The meta-analytic approach combines the results of numerous small-sample studies, thus provides overall results based on a sufficient sample.

Applications of Schmidt and Hunter's meta-analysis procedure has shown, in numerous studies, that much of the observed variation in validities for similar job-test combinations is artifactual. For example, Schmidt, Hunter & Caplan (1981) investigated the transportability of validities of four types of cognitive tests. They reported that "support for generalizability was substantiated for general mental ability and arithmetic tests" (p 261). They found that sampling error alone accounted for 90 percent of all variance due to artifacts. Additional support is given by Schmidt and Hunter (1984), Schmidt, Hunter, Pearlman & Shane (1979), and Pearlman, Schmidt & Hunter (1980).

The evidence sufficiently subdues the situational specificity hypothesis. For a given job-test combination, the four sources of error corrected for by the meta-analytic procedure are capable of producing as much variance as is generally observed between validation studies.

Validity Generalization and the ASVAB

It appears that a general ability, or g factor, serves as a link among validities of cognitive ability tests. Though authors disagree on the proper definitions of g and intelligence, they agree that differences among individuals in a general ability factor are largely responsible for differences in success and status in the United States (Tyler, 1986).

Though general cognitive ability is seen as a valuable predictor of performance, with some validity for all jobs, there are jobs for which its validity is relatively low. Data available from U.S. Employment Service (USES) validation studies, on 515 jobs representative of those in the Dictionary of Occupational Titles (DOT) point toward job complexity as a moderator of validities (Hunter, 1984). Even in the worst cases, however, Hunter concluded an average cognitive ability validity of .37 for training success and .32 for job proficiency. An additional analysis of Army data (Helme, Gibson & Brogden, 1957) indicated a higher validity for cognitive ability on complex tasks (decision making) accompanied by higher validities for psychomotor ability measures on tasks of low complexity (following instructions) (Hunter, 1984). Thus as job complexity varies, higher predictive validities may be obtained by shifting between cognitive and psychomotor abilities as the primary predictors.

If cognitive ability is such a pervasive concept, what then, is its relationship to the ASVAB, which was developed to reflect differential aptitudes? The theory of differential aptitudes supposes that ASVAB subtest composite (AI) validities will be high for jobs in corresponding occupational areas, and lower for unrelated occupations. For specific jobs, poor predictability of one aptitude could thus be offset by emphasizing another aptitude with higher validity. Hunter, Crosson & Friedman (1985) disbelieved this theory, and subsequently showed it to be untrue. For their samples of high school students who took the ASVAB, (N ranged from 596 to 13,904), each occupational composite (selector AI) was nearly as valid for other occupational areas as it was for its own. A clear lack of differential validity was observed.

In an Air Force sample of 29,619 recruits, with AI validities computed for 70 jobs, only in the electronics area was the AI more valid for its corresponding jobs than for any of the other composites (Wilbourn, Valentine & Ree, 1984). In the same study, mechanical and general occupational performances were predicted better by composites other than the ones designed specifically for these career fields. Predictive ability of a General Cognitive ability Composite was found to be as high as those for Administrative and Electronics composites, and higher than those for the other two composites (Mechanical and General).

Viewing such results as these, we can see a strong indication of a general cognitive ability factor moderating the relationships between ASVAB subtest validities. According to Jensen (1986):

> Virtually no one today disputes that a g factor can be extracted from the correlations among any large and diverse collection of mental ability tests, and that the g factor is usually substantiated in the

> sense of subsuming a relatively large proportion of the total variance in all of the tests as compared with other factors besides g. The point that is being questioned is whether the g factor represents any reality outside the operations of psychometric tests and factor analysis (p 302).

Extensive evidence does support the existence of the g phenomenon. Non-psychometric correlates of g have been identified in recent years. g loadings have been found related to heritability (Block, 1968; Tambs, Sundet & Magnus, 1984), and to test correlations among family members (Nagoshi & Johnson, 1986). A review by Jensen (1983) identified 12 studies providing evidence that g loadings are related to inbreeding depression. Other reviews (Eysenck & Barrett, 1985; Haier, Robinson, Braden & Williams, 1983) have led to the conclusion that g correlates with evoked electrical brain potentials.

Jensen (1986) went on to describe a hierarchical factor structure wherein variance unique to individual tests is filtered out at the level of primary factors, variance unique to primary factors is filtered out at the level of secondary factors, etcetera, with the g factor appearing at the highest level. He states:

> The g factor is remarkably stable across different collections of mental tests, even collections of

> tests that bear hardly any superficial resemblance to one another . . . . all so-called intelligence tests, or "IQ" tests, even when they have not been constructed with reference to factor analysis, are found to be very highly g loaded (p 308 and 310).

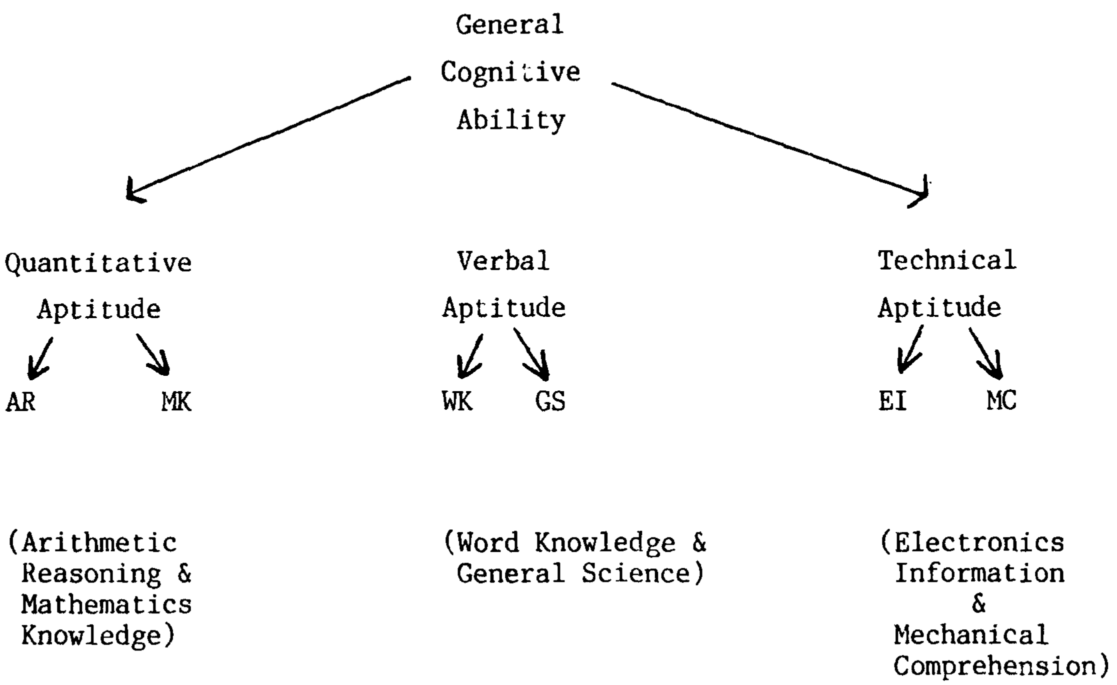
Though the ASVAB was not constructed via factor analytic methods, researchers have applied such procedures to the test to determine its factor structure. These analyses have indicated that the regression-based composites do not correspond to factor-analytic composites (Hunter, 1983b; U.S. Naval Personnel Research Activity, 1981).

In Hunter's investigation (1983b) he produced a hierarchical factor model of military subtests through analysis of available factor-analyzed data (Maier & Grafton, 1981; Sims & Hiatt, 1981; Thorndike, 1957). He identified three factors common to all the data: Quantitative, Verbal, and Technical composites. The factor model identified by Hunter (see figure 1) exemplifies the hierarchical structure described by Jensen (1986). Because Hunter's extensive study clearly describes the relationships among ASVAB subtests, it will be discussed in some detail.

Hunter (1983b) investigated additional data (Kass, Mitchell, Grafton & Wing, 1982), based on ASVAB forms 8, 9 and 10, using confirmatory factor analysis (N = 98,689). Only 6 of the 10 ASVAB subtests were used. One subtest, AS, did not load

Figure 1

Hierarchical Factor Model


General Cognitive Ability
Quantitative Aptitude
Verbal Aptitude
Technical Aptitude
AR
MK
WK
GS
EI
MC
(Arithmetic Reasoning & Mathematics Knowledge)
(Word Knowledge & General Science)
(Electronics Information & Mechanical Comprehension)

significantly on any factor. The inclusion of AS in the Technical factor with EI and MC would have generated inconsistent data. For this reason, AS was omitted from further analysis. Hunter found NO and CS (the two Perceptual Speed subtests) to correlate highly with each other, yet NO consistently correlated higher with other subtests than did CS. He found that, in the presence of good verbal and quantitative measures NO did not raise validity meaningfully. Also, NO and CS did not appear to measure the same construct as Perceptual Speed subtests of other batteries, thus these subtests were excluded for the reasons specified. The last exclusion was PC. Though very similar in composition to WK, PC had a reliability of about .75, while the reliability of WK was found to be .90. Weighting schemes for combining the two by giving more weight to WK did not improve reliability, so PC was dropped. Hunter stressed that these four subtests were not excluded for reasons of invalidity, but for the purpose of allowing construction of composites that are comparable with civilian test battery data.

Hunter found that the remaining six subtests produced three highly intercorrelated factors, designated as Quantitative Aptitude (AR + MK), Verbal Aptitude (WK + GS), and Technical Aptitude (EI + MC). His analyses showed that validity is not lower when analyses are performed without the four excluded subtests. Table 2 presents the ASVAB composite structure, and the factor structure

Table 2

ASVAB Composite and Factor Structures

						General Cognitive Ability					
				Percep Speed[2]		Quant[2,3]		Verbal[2,3]		Tech[2,3]	
ASVAB Subtests →		AS°	PC°	NO°	CS°	AR	MK	WK	GS	EI	MC
Air Force Composites ↓	AFQT		X	X		X		X			
	M	X							X		X
	A		X	X	X			X			
	G		X			X		X			
	E					X	X		X	X	

° Indicates composites excluded from the model.

[2] Indicates the four factors initially derived through exploratory factor analysis.

[3] Indicates the three factors isolated by Hunter's confirmatory factor analysis.

isolated by Hunter. He concluded that "factor analytic composites are more valid than the multiple-regression composites for predicting job performance (Hunter, 1984b, p 72).

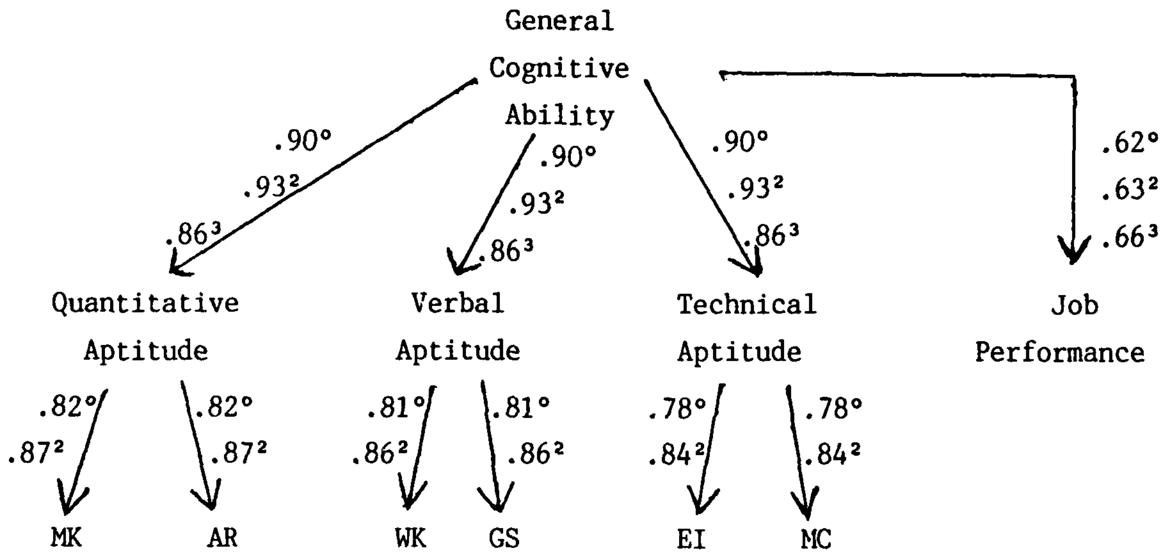
Through a subsequent confirmatory factor analysis, Hunter identified a second-order factor which he labeled General Cognitive Ability. The best estimate of this general factor was defined as the sum of the three aptitude composites (Quantitative + Verbal + Technical), or as the sum of the six underlying subtests. Table 2 shows the composition of General Cognitive Ability.

In yet another confirmatory factor analysis from the same study, Hunter analyzed the data available from Sims and Hiatt (1981), and Maier and Grafton (1981), both derived from ASVAB forms 8, 9 and 10, and he also included data from Thorndike (1957) based on the now obsolete Army AC-1B. The Maier and Grafton, and Sims and Hiatt data produced the path model shown in figure 2. The Thorndike data produced a similar pattern, but with two changes relative to the model. An additional primary factor, Perceptual Aptitude, was obtained consisting of now obsolete subtests. Also, two of the AC-1B primary factors contain subtest configurations different from the other data (due to test revisions over the years), thus these subtest validities are not shown in figure 2.

That all data sources produced a similar pattern of aptitudes provides strong support for existence of the second-order factor, General Cognitive Ability. Results from all three data sources

Figure 2

Path Model relative to military subtest data.



[°] Sims and Hiatt (1981) data.

[2] Maier and Grafton (1981) data.

[3] Thorndike (1957) data.

indicate that the correlations between the aptitude factors differ by no more than sampling error (see Table 3). Hunter's measurement model provides evidence that the individual ASVAB subtests contribute to overall validity via their contributions to one of the three primary aptitudes. The validities of the primary aptitudes, in turn, are not indicative due to independent contributions, but to their inclusion in the second-order factor, General Cognitive Ability. The general factor provides a measure of job performance due to the mediating effects of job knowledge.

General Cognitive Ability was seen to predict job training performance in specific areas as well or better than the regression-based composite of specific abilities designed expressly for that occupational area. Only one exception was observed. Perceptual Speed made an incremental contribution to validity of .02 over General Cognitive Ability, in predicting performance for clerical occupations (Hunter, 1984b). Hunter presumed this was due to effects of the speeded ASVAB subtest, CS. His (1985) meta-analysis of eight military studies verified the evidence for differential validity in business and clerical occupations, due to the Perceptual Speed factor. The validity of Speed for clerical work is .43, versus .37 for other types of work. Hunter identified Perceptual Aptitude as improving validities beyond the influence of General Cognitive Ability, relative to Thorndike's data (1957). Current versions of the ASVAB however, do not contain Perceptual

Table 3

Intercorrelations of factors.

Aptitude				
Factors	Q[°]	V[2]	T[3]	G[*]
Q	1.00			
V	.82	1.00		
T	.80	.89	1.00	
G	.86	.96	.93	1.00

[°] Q (Quantitative Aptitude) = AR + MK

[2] V (Verbal Aptitude) = WK + GS

[3] T (Technical Aptitude) = EI + MC

[*] G (General Cognitive Ability)

Aptitude measures.

Hunter's extensive work on military data has shown the ASVAB subtest composites to be of acceptable validity, though in light of General Cognitive Ability as a moderating factor, of perhaps questionable utility.

Meta-Analysis

Though Hunter and Schmidt's validity generalization procedure has been shown to be a viable procedure, it has not been without critics. Forty questions regarding meta-analysis and its application to validity generalization are presented in a debate format (Sackett, Schmitt, Tenopyr, Kehoe & Zedeck, 1985; Schmidt, Hunter, Pearlman & Hirsch, 1985). For 23 of the questions, Schmidt, et al.'s answers stand as acceptable. Schmidt (1988) comments:

> The commentators . . . . took no issue with the major practical conclusion of meta-analytic research in personnel selection: that validities, particularly of cognitive tests, have been shown to be widely generalizable across settings, jobs, populations, organizations, geographical areas, time periods, etc. Most of the commentary was in the nature of attempts to "fine tune" statistical methods or applications (p 179).

Practical acceptance of Schmidt and Hunter's meta-analysis model is evident from the wide application it has been given, including topics outside the area of personnel selection. Examples include: correlates of role conflict and role ambiguity (Fisher & Gittelson, 1983), evaluation of Fiedler's theory of leadership (Premack & Wanous, 1985), and ability of financial analysts to predict stock growth (Coggin & Hunter, 1983).

The USES and the federal government are currently the two largest users of validity generalization. The USES has adopted validity generalization as the basis for its testing program that operates through state employment services. Civilian users include the American Petroleum Institute, Sears-Roebuck, and various insurance and utilities industries (Schmidt, 1986).

Validity generalization concepts have been incorporated in psychological texts (Anastasi, 1982), and in the Standards for Educational and Psychological Testing. The Uniform Guidelines on Employee Selection Procedures issued in 1978 are more accepting of validity generalization than the previous (1970) guidelines, and "spell out more precisely the methods necessary to "borrow validity" or generalize validity results to similar jobs" (Baker & Terpstra, 1982, p 603). Baker & Terpstra specify two court cases in which generalized validities have been accepted as evidence to refute racial and sex discrimination charges: _Friend v. Leidinger_, 1977, and _Pegues v. Mississippi State Employment_

Service, 1980. It is thus not unreasonable to expect that future revisions of the guidelines will include an even broader acceptance of validity generalization procedures.

Statement of Purpose

The present study investigates characteristics of ASVAB validity data via Hunter and Schmidt's meta-analysis procedures. In light of the research reviewed, it follows that General Cognitive Ability moderates the relationship between job knowledge and job performance. Also, verbal and numerical skills constitute the major components of General Cognitive Ability. The AFQT, containing two verbal subtests (WK and PC), plus two numerical subtests (AR and NO), may be considered to represent a general cognitive measure.

Hunter's work in particular (1983a, 1983b, 1984, 1985) has demonstrated higher predictive validities for General Cognitive Ability than for the regression-based ASVAB selector composites. This study extends Hunter's work by attempting to demonstrate that the AFQT, as a general cognitive measure intrinsic to the ASVAB, will prove to be a valid predictor of Air Force training course success. In line with Hunter's results, the general cognitive measure (AFQT) expected to improve prediction beyond that given by the four selector AIs (Mechanical, General, Administrative, and Electronic). Furthermore, it is expected that the improvement in prediction will be greater for women that for men. This is expected for two reasons. First, females have been shown to score somewhat higher than men on cognitive ability measures, thus the female's validities should raise when the AFQT alone is used as a predictor. Secondly, general cognitive measures (as with the AFQT) lack mechanically-related items, on which men tend to score higher than women.

The formal hypotheses tested are:

1. $\rho_{AFQT} = 0$

   $\rho_{AFQT} > 0$

2. $\rho_{AFQT} = \rho_{M}$ for Mechanical AI

   $\rho_{AFQT} > \rho_{M}$

3. $\rho_{AFQT} = \rho_{G}$ for General AI

   $\rho_{AFQT} > \rho_{G}$

4. $\rho_{AFQT} = \rho_{A}$ for Administrative AI

   $\rho_{AFQT} > \rho_{A}$

5. $\rho_{AFQT} = \rho_{E}$ for Electronic AI

   $\rho_{AFQT} > \rho_{E}$

6. $\rho_{AFQT.F} - \rho_{AIs.F} = \rho_{AFQT.M} - \rho_{AIs.M}$

   $\rho_{AFQT.F} - \rho_{AIs.F} > \rho_{AFQT.} - \rho_{AIs.M}$

CHAPTER 2

METHODS

The Instrument

Current ASVAB versions (forms through 14) consist of 334 items in 10 subtests. The ASVAB is group administered by trained DoD personnel. Administration sites are nation-wide, in Military Entrance Processing Stations, and in high schools.

ASVAB subject areas were chosen according to their validity for predicting training criteria in each of the military services. The content of subtests of forms 8 through 14 is shown in Table A-1. The total testing time is 144 minutes.

Subjects

Data were obtained from an AFHRL validation study performed by Wilbourn, Valentine & Ree (1984). The subjects for their research consisted of 29,619 males and females, aged 17 to 24. A breakdown of these subjects is given in Table A-2. The subjects were all first-term Air Force enlistees, most of whom were high school graduates. They were tested on ASVAB forms 8, 9 and 10 between October, 1980 and March, 1982.

The subject group was restricted in range to the extent that only those passing certain cutoff scores gained enlistment. Selection criteria were: AFQT score; the sum of the four AIs;

and the General AI score. An additional restriction factor curtailed the technical school samples at the upper score levels. Because each technical school maintains a cutoff score on an appropriate selector AI, higher-scoring subjects were assigned to schools with higher eligibility requirements.

The Criterion

The validation criterion used was technical training course final grades. The courses included in the study were the 70 courses which had at least 100 graduates. Technical school grades generally range between 70 (passing) and 100. Approximately four percent of enlistees do not pass the training course. Attrition thus further restricted the sample range. As previously discussed, training school success is a highly appropriate criterion, as it directly predicts job performance (Hunter, 1984).

Procedures

Meta-analytic procedures will be used to combine validities given in the Wilbourn, et al. data. Reliability measures will be obtained from another recent AFHRL study (Palmer, Hartke, Ree, Welsh & Valentine, 1988). Though Hunter and Schmidt have provided formulae for correction of reliabilities, it was decided that reliabilities for this study would not be corrected. Since the ASVAB is an on-line operational test, the reliability values as given will prove more informative than would corrected values.

Procedural steps are as follows:

1. Correct validity coefficients for the effects of restriction in range;
2. correct for sampling error by weighting corrected validities by the N of the appropriate sample, utilizing Fisher's z transformation;
3. estimate the mean and standard deviation for each condition; then
4. determine the significance of differences between variables.

For step four, no statistical significance test is applicable, since there is no known method for determining the standard error of a corrected correlation. Thus the method for this study will be to build confidence intervals around corrected validities, utililzing a 95 percent, or 2-standard deviation confidence interval as the decision point.

Computations will be performed via computer analysis. Formulae are given in Appendix A.

Table A-1

ASVAB Subtest Content

Subtest	Content	Number of questions
General Science (GS)	general science, including biology and physics	25
Arithmetic Rea3oning[2] (AS)	arithmetic word problems	30
Word Knowledge[2] (WK)	selecting synonyms	35
Paragraph Comprehension[2] (PC)	ability to understand a written text	15
Numerical Operations[2 3] (NO)	simple, arithmetic calculations	50
Coding Speed[3] (CS)	substituting numeric codes for verbal material	84
Auto and Shop Information (AS)	automobile and tool-usage knowledge	25

(Table continues)

Subtest	Content	Number of questions
Mathematics Knowledge (MK)	calculations including algebra, geometry, and elementary trigonometry	25
Mechanical Comprehension (MC)	general mechanical and physical principles	25
Electronics Information (EI)	electrical principles and terminology	20

[2] AFQT subtests.

[3] Speeded subtests.

Table A-2

Subject Breakdown.

	N	Percent
White males	21,554	72.8
White females	2,702	9.1
Black males	4,040	13.6
Black females	590	2.0
Other race	733	2.5
Totals	29,619	100.
Total white	24,256	81.9
Total black	4,630	15.6
Other race	733	2.5
Totals	29,619	100.
Total male	26,259	88.7
Total female	3,360	11.3
Totals	29,619	100.

REFERENCES

Anastasi A. (1982). Psychological Testing. (5th Ed.) New York: Macmillan.

Baker, D. D. & Terpstra, D. E. (1982). Employee selection: must every job be validated? Personnel Journal, 61, 602-605.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannon, R. (1978). Testing for fairness with a moderated multiple regression strategy: an alternative to differential analysis. Personnel Psychology, 31, 233-241.

Block, J. B. (1968). Hereditary components in the performance of twins on the WAIS. In S. G. Vandenberg (Ed.), Progress in Human Behavior Genetics. Baltimore: The Johns Hopkins University Press.

Bock, R. D. & Mislevy, R. J. (1981). Data Quality Analysis of the Armed Services Vocational Aptitude Battery. Chicago, IL: National Opinion Research Center.

Coggin, T. D. & Hunter, J. E. (1983). Problems in measuring the quality of investment information: the perils of the information coefficient. Financial Analysts Journal, May/June, 1-10.

Eysenck, H. J. & Barrett, P. (1985). Psychophysiology and the measurement of intelligence. In C. R. Reynolds & V. Wilson (Eds), Methodological and Statistical Advances in the Study of Individual Differences. New York: Plenum.

Fisher, C. D. & Gittelson, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. Journal of Applied Psychology, 68, 320-333.

Friedman, D., Streicher, A., Wing, H., Grafton, F., & Mitchell, K. (1983). Reliability of scores for fiscal year 1981 Army applicants: Armed Services Vocational Aptitude Battery forms 8, 9, and 10. U.S. Army Research Institute for the Behavioral and Social Sciences. Alexandria, VA.

Ghiselli, E. E. (1966). The Validity of Occupational Aptitude Tests. New York: John Wiley & Sons.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. V. (1977). Integrating findings: the meta-analysis of research. Review of Research in Education, 5, 351-379.

Gottfredson, L. S. (1986). Societal consequences of the g factor in employment. Journal of Vocational Behavior, 29, 379-410.

Graham, J. R. & Lilly, R. S. (1984). Psychological Testing. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Haier, R. J., Robinson, D. L., Braden, W., & Williams, D. (1983). Electrical potentials of the cerebral cortex and psychometric intelligence. Personality and Individual Differences, 4, 591-599.

Hawk, J. (1986). Real world implications of g. Journal of Vocational Behavior, 29, 411-414.

Helme, W. E., Gibson, W. A., & Brogden, H. E. (1957). An empirical test of shrinkage problems in personnel classification research. Personnel Board Technical Research Note 84.

Hunter, J. E. (1983a). The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of the general factors over specific factors in the predictions of job performance for USES. Test Research Report No. 44, U.S. Department of Labor, U.S. Employment Services, Washington, DC.

Hunter, J. E. (1983b). Validity generalization of the ASVAB: higher validity for factor analytic composites. Research Applications, Inc., Rockville, MD.

Hunter, J. E. (1984). The validity of the ASVAB as a predictor of civilian job performance. Research Applications, Inc., Rockville, MD.

Hunter, J. E. (1985). Validity generalization of the ASVAB: preliminary report. Research Applications, Inc., Rockville, MD.

Hunter, J. E., Crosson, J. J., & Friedman, D. H. (1985). The validity of the Armed Services Vocational Aptitude Battery (ASVAB) for civilian and military job performance. Research Applications, Inc., Rockville, MD.

Hunter, J. E. & Schmidt, F. L. (1982). Ability tests: Economic benefits versus the issue of fairness. Industrial Relations, 21, 293-308.

Hunter, J. E. & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. American Psychologist, 38, 473-478.

Hunter, J. E., Schmidt, F. L. & Hunter, R. (1979). Differential validity of employment tests by race: a comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Jensen, A. R. (1980). Bias in Mental Testing. New York: Macmillan Publishing Co., Inc.

Jensen, A. R. (1983). Effects of inbreeding on mental ability factors. Personality and Individual Differences, 4, 71-87.

Jensen, A. R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. The Behavioral and Brain Sciences, 8, 193-219.

Jensen, A. R. (1986). g: artifact or reality? Journal of Vocational Behavior, 29, 310-331.

Kass, R. A., Mitchell, K., Grafton, F., & Wing, H. (1982). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10: Army applicant sample. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Maccoby, E. E. (Ed.) (1966). The Development of Sex Differences. Stanford, CA: Stanford University Press.

Maccoby, E. E., & Jacklin, C. N. (1974). The Psychology of Sex

Differences. Stanford, CA: Stanford University Press.

Maier, M. H., & Grafton, F. C. (1981). *Aptitude composites for ASVAB forms 8, 9, and 10*. Research Report 1308, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

McCarthy, D. (1954). Language development in children. In L. Carmichael (Ed.) Manual of Child Psychology, 2nd Ed., New York: Wiley.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist*, 19, 871-882.

Nairn, A. & Associates (1980). *The reign of ETS: The corporation that makes up minds*. Washington, DC: Nader.

Nagoshi, C. T., & Johnson, R. C. (1986). The ubiquity of g. *Personality and Individual Differences*, *7*, 201-207.

Palmer, P., Hartke, D. D., Ree, M. J., Welsh, J. R. & Valentine, L. D. Jr. (1988). *Armed Services Vocational Aptitude Battery (ASVAB): Alternate form reliability (forms 8, 9, 10, and 11)*. Report No. AFHRL-TP-87-48, Air Force Human Resources Laboratory, Brooks Air Force Base, TX.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, *65*, 373-406.

Premack, S., & Wanous, J. P. (1985). Meta-analysis of realistic job preview experiments. *Journal of Applied Psychology*, *70*, 706-719.

Ree, M. J., Mathews, J. J., Mullins, C. J., & Massey, R. H. (1982). *Calibration of Armed Services Vocational Aptitude Battery forms 8, 9, and 10*. Report No. AFHRL-TR-81-49, Air Force Human Resources Laboratory, Brooks Air Force Base, TX.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*, 1-62.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, *85*, 185-193.

Sackett, P. R., Schmitt, N., Tenopyr, M. L., Kehoe, J., & Zedeck, S. (1985). Commentary on forty questions about validity generalization and meta-analysis. *Personnel Psychology*, *38*, 697-798.

Sims, W. H., & Hiatt, C. M. (1981). *Validation of the Armed Services Vocational Aptitude Battery (ASVAB) forms 6 and 7 with application to ASVAB forms 8, 9, and 10*. Marine Corps Operations Analysis Group, CNS 1160, Center for Navy Analysis, Alexandria, VA.

Schmidt, F. L. (1988). Validity generalization and the future of criterion-related validity. In H. Wainer & H. I. Braun (Eds.) *Test Validity*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of*

Applied Psychology, 62, 529-540.

Schmidt, F. L. & Hunter, J. E. (1978). Moderator research and the law of small numbers. Personnel Psychology, 31, 215-232.

Schmidt, F. L., & Hunter, J. E. (1984). A within setting empirical test of the situational specificity hypothesis in personnel selection. Personnel Psychology, 37, 317-326.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 66, 261-273.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 66, 166-185.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validity studies. Journal of Applied Psychology, 61, 473-485.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The ecomonic impact of job selection methods on size, productivity, and payroll costs of the federal work force: an empirically based demonstration. Personnel Psychology, 39, 1-29.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsch, H. R. (1985). Forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697-798.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 33, 705-724.

Smith, I. M. (1948). Measurement of spatial ability in school pupils. Occupational Psychology, 22, 150-159.

Spearman, C. (1904). "General intelligence", objectively determined and measured. American Journal of Psychology, 15, 201-293.

Spearman, C. (1927). The Abilities of Man. London: McMillan.

Tambs, K., Sundet, J. M., & Magnus, P. (1984). Heritability analysis of the WAIS subtests. A study of twins. Intelligence, 8, 283-293.

Thorndike, R. L. (1957). The optimum test composites to predict a set of criteria. Reoort No. AFPTRC-TN-57-103, Air Force Personnel Training and Research Center, Lackland Air Force Base, TX.

Thurstone, L. L. (1938). Primary Mental Abilities. Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1941) Factorial Studies of Intelligence. Chicago: University of Chicago Press.

Tyler, L. E. (1986). Back to Spearman? Journal of Vocational Behavior, 29, 445-450.

U.S. Department of Defense (1984). Armed Services Vocational Aptitude Battery (ASVAB): Test Manual. North Chicago, IL: U.S. Military

Entrance Processing Command.

U.S. Naval Personnel Research Activity. (1981). Basic Test Battery Validity Report for 89 Class "A" School Samples. San Diego, CA.

Wilbourn, J. M., Valentine, L. D. Jr., & Ree, M. J. (1984). Relationships of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10 to Air Force technical school final grades. Report No. AFHRL-TP-84-8. Air Force Human Resources Laboratory, Brooks Air Force Base, TX.

Additional References to be included in final copy:

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing. Old theories and new research findings. American Psychologist, 36, 1128-1137.

APPENDIX A

COMPUTATIONAL FORMULAE FOR

META-ANALYTIC PROCEDURES

1. Correction for restriction in range:

$$U = \frac{s}{S}$$

Where: s = standard deviation of the independent variable (ASVAB scores)

and: S = standard deviation of the dependent variable (final school grades)

$$\rho_1 = \frac{U \rho_2}{\sqrt{(U^2 - 1)\rho_2^2 + 1}}$$

Where: ρ_1 = reference population correlation

and: ρ_2 = study (sample) correlation

2. Correction for sampling error:

$$\sigma_e^2 = \frac{(1 - \bar{r}^2)^2 K}{N}$$

Where: K = the number of studies or values

and: $N = N_i$ (total sample size)

3. Estimating mean and variance:

Mean: $$\bar{r} = \frac{\Sigma (N_i r_i)}{\Sigma N_i}$$

Corresponding variance: $$s^2{}_r = \frac{\Sigma [N_i (r_i - \bar{r})^2]}{\Sigma N_i}$$

4. Confidence intervals:

Step 1: Build confidence intervals around corrected correlations.

Step 2: Use formula for range restriction to correct endpoints of the confidence interval.

$$r_c = \frac{U\rho_2}{\sqrt{(U^2 - 1)\rho_2^2 + 1}}$$

VITA

CENSUS:

Nicole S. Stermer [redacted]. Her parents are [redacted], of Dickinson, North Dakota. She is married with no children.

TRAINING:

Nicole S. Stermer graduated from Dickinson High School, Dickinson, North Dakota, May, 1976. She received her bachelor of arts degree from the University of North Dakota, Grand Forks, North Dakota, December, 1982. She has performed graduate work in 1984 at Georgia College, Milledgeville, Georgia, and in 1986 and 1987 at the University of Texas at San Antonio, Texas.

EXPERIENCE:

Publications and similar achievements:

In 1982, she completed an undergraduate thesis in psychology. From 1983 until 1985 she taught mentally retarded adults for the Houston County Association for Exceptional Children, Warner Robins, Georgia. In 1985 she was commissioned in the U. S. Air Force and presently serves as a Behavioral Scientist.

[redacted] [redacted]

[redacted] [redacted]